

面向隐私保护的可视分析

一、隐私保护

1. 问题描述

在这个信息量爆炸增长的时代，数据中蕴含的价值吸引了越来越多的目光。小到餐馆定价，大到政府调控，有了数据的支持，决策变得更加容易且有针对性。大数据的观念已经逐渐渗透到各行各业中，并日益积蓄着力量，以为人们提供更好的服务。与此同时，为了支持这些分析、挖掘需求，数据收集变得无处不在，从城市中的传感器到手机上的 APP，我们的信息也在不经意间被记录下来。一条条数据汇聚在一起，推动着社会的发展，也为窥探隐私的不法分子营造了难得的“机遇”。

信息的暴露会给人们的生活带来困扰，也会对人身和财产安全造成威胁。我国法律保护人们的隐私权，其中包括保护个人信息不为他人知悉。但是，为了寻求一些服务，我们不可避免地需要将一些信息与他人共享。比如说在使用实时路线规划功能，或者了解周边的美食时，打开 GPS 定位功能是必要操作。这时，我们的位置信息就会被上传。为了更好地使用数据，保护数据中的隐私就成了亟待解决的重要问题。

2. 保护对象

在个体向一些组织或机构寻求服务时，可能需要登记个人信息。此外，一些组织或机构也会以调查等方式收集个体数据。虽然在分析数据时使用的往往是由个体数据会聚而成的群体数据，但是毫无疑问，群体数据中是包含个人信息的。本段将从数据类型的角度，对隐私保护的保护对象进行描述。

2.1 表格数据

全部患者的体检报告、员工的业绩表等就是典型的表格数据。在这类数据中，个体以数据行的形式存在，表格的列则对应多种与个人相关的属性信息，如性别、年龄等。这些属性信息中，既存在着与潜在分析任务密切相关的信息，也存在着相对私密的信息，如患病情况、具体收入等。属性如姓名、身份证号等会直接暴露个体身份，且对分析没有帮助的信息可以直接从数据中去除。然而，一些与分析直接或间接相关的属性若被去除，会导致科研人员无法从中分析出有效信息。举个例子，医学领域的研究人员需要病患数据来完善预防、控制疾病的措施。此时病人的生活习惯、疾病等信息都是研究必要的。同时，这些信息也是一些病患不愿让他人知道的。此时就需要有效的手段来解决隐私保护和数据分析之间的矛盾了。

2.2 图数据

图数据也被称为关系数据。这种数据包含节点和链接两部分。其中，节点对应的就是个体，链接则对应着节点之间的某种关系。这种关系在不同图数据中往往有着不同的语义，在复杂的图数据中，也可能同时存在代表不同语义的边。比如，在社交关系数据中，链接可能对应着好友关系；在交易网络中，链接代表着一次或多次交易。这些链接将原本独立的个体组合起来，构成了图结构。我们可以从图结构中获取到的信息不止是节点和链接的对应关系，还有它们的结构特征，比如度数，从属社团等。结构特征的加入使图数据较表格数据而言更加复杂，相应的隐私保护方法也需要考虑更多细节。

2.3 轨迹数据

轨迹数据由一串时间及相应的位置组成。根据数据的采集频率及方式的不同，轨迹数据的质量存在较大差异。一些需要精确定位的系统如地图 APP 可以记录个体的准确位置点。而像基站等区域性服务则只能对个体所在的区域进行记录。另外，一些卡口数据虽然可以记录精确的轨迹位置，但只有个体在经过固定的出入口时，信息才能被记录。诸如此类的原因限制了轨迹数据的精确度。但是轨迹信息泄露的危害性不容低估。轨迹信息，特别是周期性

轨迹信息（如每天早起去公园锻炼身体）被不法分子得知，可能会影响个体的正常生活。

3. 隐私暴露的方式

隐私暴露的方式主要有两种。一种为身份泄露，即攻击者可以识别数据中个体的身份，从而了解该个体的全部信息。另一种为信息泄露。在这种情况下，攻击者虽然无法辨识个体的身份，但却可以结合自己对个体的了解，基于数据中的信息推断出个体的敏感信息内容。

3.1 身份泄露

可以用来识别个体身份的信息不止是姓名、身份证号等可以唯一标识身份的信息。其他信息通过叠加也有可能暴露身份信息。以一个班级的学生的兴趣爱好为例，喜欢打篮球的学生可能有很多，喜欢书法的同学也有几个，但是既喜欢打篮球，又喜欢书法的同学可能只有一个。当信息量增加时，通过多种信息识别个体的概率就会大大提高。当身份信息暴露时，数据库中的所有关于个体的信息都将会与个体对应起来，被攻击者知晓。

3.2 信息泄露

有的时候，攻击者无法通过数据对个体的身份进行确认，但是，基于数据中信息的整体分布，以及对个体其他信息的了解，攻击者可能通过一些分类方法对个体的敏感信息进行推测，从而以一定概率知晓敏感信息的正确内容。这种情况显然是人们不想看到的。我们保护隐私的目的就是为了阻止个体的敏感信息被他人探知。信息泄露对应的就是这种情况。预防信息泄露的要求明显比阻止身份泄露的难度要大。考虑到分析可能会用到数据的整体分布信息，在保护隐私的前提下尽量维护数据的实用性就更加困难。

4. 常见方法

Dasgupta 和 Kosara [1] 对可行的隐私保护方法进行了总结（见图 1）。他们认为，从原始数据出发，经过不同的处理，可以通过视觉聚类 and 数据聚类两种手段实现具有保障性的隐私保护。实际上，隐私保护的方法不局限于聚类方法。在本段中，我们将从视觉空间和数据空间两个方面来对常见的隐私保护方法进行介绍。

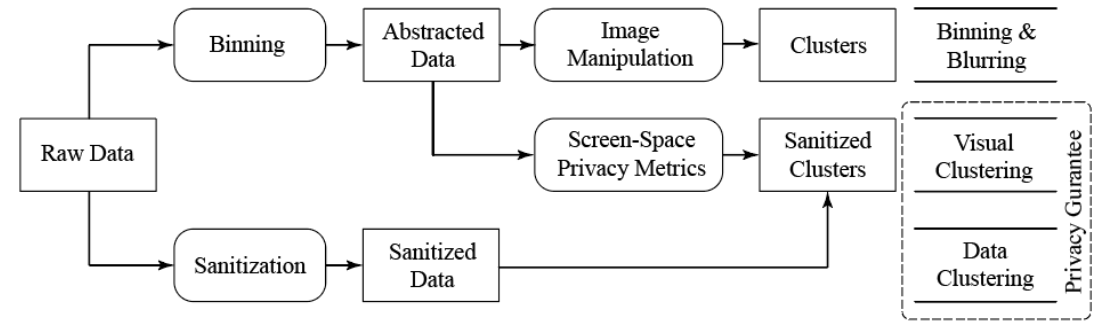


图 1: Dasgupta 和 Kosara 总结的可行的隐私保护方法[1]

4.1 视觉空间

在视觉空间对数据进行处理，即对数据的可视表达进行限制。这是一种针对使用可视分析的分析人员提供的隐私保护方法。在可视分析中，可视化向分析人员基于视觉通道传达信息。不同的可视表达可以对信息给出不同的诠释。当可视表达中存在遮挡、模糊等问题时，观看可视化的人就会对数据的具体情况产生不确定性。这种不确定性可能会在一定程度上干扰分析，但合理运用也可以实现隐私保护的效果。

4.2 数据空间

从数据空间入手来处理数据的隐私保护方法，允许分析人员通过一定渠道获取处理后的数据，并使用任意方式分析。本章将对两类经典模型——语义匿名模型和差分隐私模型给出介绍。

4.2.1 语义匿名模型

语义匿名模型指通过模糊信息等方式,使个体和个体之间的信息从数据中无法区分。其中,无法被攻击者区分的数据集合被称为等价类。最早的语义匿名模型是由 Sweeney 提出的 k -anonymity 模型[2]。该模型要求每个等价类中的数据项个数不少于可定义参数 k ,即每个个体至少和其他 $k-1$ 个个体不可区分。为了实现这一目的,具体的数据处理方法为将等价类进行合并。举个例子,假如数据库中年龄为 0~9 岁的个体有一人,10~18 岁的个体各有两人。若用户将 k 设定为 2,那么,0~9 岁的等价类就不满足隐私保护需求。为了使每个等价类中的个体数量超过 2,可以将他们的年龄都显示为 0~18 岁,就可以形成一个大的等价类。在新的等价类中共有 3 个个体,也就满足了要求。可以看出,该方法对个体的身份实现了比较好的匿名。

在此之后,考虑到即使身份没有泄露,也存在信息泄露的可能性, l -diversity [3]和 t -closeness [4]等方法也被相继提出。这些模型对等价类提出了更高的要求,即等价类中的敏感信息种类需要有足够的差异性,以及分布需要和整体数据分布近似。对于此类方法来说,参数值的设定决定了隐私保护级别的高低。级别越高,对等价类的要求越高,需要进行合并的等价类越多,对数据质量的影响越大。

4.2.2 差分隐私模型

差分隐私模型[5]是基于随机算法对信息添加扰动,从而使攻击者对整个数据集的任一信息都不能完全确定,以实现隐私保护的目。差分隐私的具体定义如下:

一个随机函数 K 可以提供 ϵ -差分隐私,当它对所有最多差一个元素的数据集 D_1 和 D_2 ,和全部 $S \subseteq \text{Range}(K)$,有

$$\Pr[K(D_1) \in S] \leq \exp(\epsilon) \times \Pr[K(D_2) \in S].$$

二、挑战

隐私保护程度的检测

(如何定义被保护;攻击模型的多样化)

与实用性之间的冲突

(冲突不可避免;实用性难以概括、量化)

三、可视化与数据感知(对应视觉空间方法)

可视表达

(用视觉编码表示数据)

不确定性

(视觉感知存在不确定性,可用于隐私保护)

四、可视分析与隐私保护流程(对应数据空间方法)

交互定制

(保护方法;参数定义)

解释过程

(复杂模型的可视解释)

参考文献

- [1] Dasgupta A, Kosara R. Adaptive privacy-preserving visualization using parallel coordinates[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2241-2248.
- [2] L. Sweeney. k -Anonymity: A Model for Protecting Privacy. IEEE Security And Privacy, 10(5):1-14, 2002.
- [3] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l -diversity: Privacy

beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):3, 2007.

- [4] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the IEEE 23rd International Conference on Data Engineering, pp. 106–115. IEEE, 2007.

- [5] Dwork C. Differential Privacy: A Survey of Results[J].